



## Durham Research Online

---

### Deposited in DRO:

18 February 2016

### Version of attached file:

Published Version

### Peer-review status of attached file:

Peer-reviewed

### Citation for published item:

Wilkinson, S. (2014) 'Accounting for the phenomenology and varieties of auditory verbal hallucination within a predictive processing framework.', *Consciousness and cognition*, 30 . pp. 142-155.

### Further information on publisher's website:

<http://dx.doi.org/10.1016/j.concog.2014.09.002>

### Publisher's copyright statement:

© 2014 The Author. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/3.0/>).

### Additional information:

---

### Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a [link](#) is made to the metadata record in DRO
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the [full DRO policy](#) for further details.



## Review

## Accounting for the phenomenology and varieties of auditory verbal hallucination within a predictive processing framework



Sam Wilkinson\*

Department of Philosophy, Durham University, 50 Old Elvet, Durham DH1 3HN, UK

## ARTICLE INFO

## Article history:

Received 26 March 2014

Available online 3 October 2014

## Keywords:

Auditory-verbal hallucination

Psychosis

Schizophrenia

Predictive processing

## ABSTRACT

Two challenges that face popular self-monitoring theories (SMTs) of auditory verbal hallucination (AVH) are that they cannot account for the auditory phenomenology of AVHs and that they cannot account for their variety. In this paper I show that both challenges can be met by adopting a predictive processing framework (PPF), and by viewing AVHs as arising from abnormalities in predictive processing. I show how, within the PPF, both the auditory phenomenology of AVHs, and three subtypes of AVH, can be accounted for.

© 2014 The Author. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/3.0/>).

## Contents

0. Introduction .....	143
1. Self-monitoring theories .....	143
1.1. Introducing self-monitoring theories (SMTs) .....	143
1.2. Support for SMTs .....	144
1.3. Applying SMTs to AVH .....	144
2. Two challenges facing the application of SMTs to AVH .....	144
2.1. The Auditory Phenomenology Challenge .....	145
2.2. The Varieties of AVHs Challenge .....	145
3. The predictive processing framework (PPF) .....	145
3.1. An initial presentation: "inference" and efficiency .....	146
3.2. Precision weighting, second-order prediction and "attention" .....	146
3.3. Conscious effects of predictive processing .....	147
3.3.1. Binocular rivalry .....	147
3.3.2. The Hollow Mask Illusion .....	147
3.3.3. The McGurk Effect .....	147
3.4. Clearing up potential confusion: what the brain does and what the person does .....	148
4. Predictive processing and psychosis .....	148
4.1. Conscious effects are altered in patients with diagnoses of schizophrenia .....	148
4.2. Evidence from eye tracking .....	148
5. Addressing the two challenges .....	149
5.1. Dissolving the Auditory Phenomenology Challenge .....	149
5.2. Accounting for three subtypes of AVH .....	150
5.2.1. Inner speech hallucinations .....	150

\* Fax: +44 191 334 6551.

E-mail address: [sam.wilkinson@durham.ac.uk](mailto:sam.wilkinson@durham.ac.uk)

5.2.2.	Memory-based hallucinations .....	150
5.2.3.	Hypervigilance Hallucinations .....	151
5.3.	Multiple models within one framework: personal history and attentional focus .....	152
6.	Summary and conclusion .....	153
	Acknowledgments .....	153
	References .....	153

## 0. Introduction

The positive symptoms of schizophrenia include delusions of control (“somebody else is controlling my actions”), thought insertion (“somebody is putting their thoughts into my head”) and auditory verbal hallucination (AVH) (hearing voices in the absence of a speaker). Perhaps the most popular theories for understanding these disparate symptoms are self-monitoring theories, which attempt to explain them as the product of one abnormality, namely, a problem with self-monitoring. According to these theories, our nervous systems distinguish self-generated from externally generated stimuli, through a process of self-monitoring. When this monitoring goes awry, self-generated stimuli are erroneously attributed to an external cause. The various positive symptoms all involve faulty monitoring and simply differ insofar as that which is failing to be properly monitored differs. In delusions of control it is bodily action, whereas in AVH and thought insertion it is widely thought to be inner speech (Feinberg, 1978; Frith, 1992; Jones & Fernyhough, 2007; Seal, Aleman, & McGuire, 2004). Although ingenious, the breadth of application of SMTs has recently been questioned by several theorists (Gallagher, 2004; Jones, 2010; Stephens & Graham, 2000; Wu, 2012). Such criticisms tend not to take issue with the application of SMTs to symptoms involving bodily action, like delusions of control and (their merely experiential analogue) illusions of passivity. Rather, they claim that SMTs struggle to account for AVH and thought insertion. In this paper, I will focus on AVH, and on two challenges in particular. They are:

*The Auditory Phenomenology Challenge* – How do you explain the auditory phenomenology of AVH if it is misattributed inner speech?

*The Varieties of AVH Challenge* – How do you account for the varieties of AVH if it is (always) misattributed inner speech?

In this paper, I suggest that both challenges can be met if we adopt a recently popular general framework for thinking about what the brain does (e.g. Clark, 2013; Friston, 2005, 2010; Hohwy, 2013) which we could call the predictive processing framework (PPF).

It is worth mentioning that the application of predictive processing to psychosis is not new. Indeed, Chris Frith, perhaps the best-known proponent of SMTs, has suggested something along these lines in Fletcher and Frith (2009). Since then, Adams, Shipp, and Friston (2013) have also suggested accounts of psychosis within the PPF. This work, however, does not focus on AVHs to the extent that I do, nor does it focus on the two challenges that I address here.

I proceed as follows. I start by presenting SMTs and show why they have been found attractive and plausible. I present the two challenges facing the application of SMTs to AVHs. I then introduce, motivate and clarify the PPF. I then present evidence suggesting that predictive processing might be disrupted in psychosis. Finally, I end by applying the PPF to voice-hearing, and show how it can, first, address the auditory phenomenology challenge, and second, nicely account for the three subtypes of AVH I present.

## 1. Self-monitoring theories

In this section I characterise SMTs, and describe the evidence that has been used to support them.

### 1.1. Introducing self-monitoring theories (SMTs)

Perhaps the first theorist to make use of self-monitoring was Helmholtz (1866). His concern, however, was not with psychopathology, but with the following problem presented by healthy visual cognition. When an image moves across the retina, how does our brain know whether it is the world moving across our eyes or our eyes moving across the world? Helmholtz suggested that our brain can tell the difference because when our eyes move there is a motor command. More specifically, information about the motor command, which Sperry (1950) later dubbed the “corollary discharge”, is used by the brain to predict the sensory consequences that would be produced by the eye movement. If the predicted and actual sensory consequences match then the brain infers that the change was self-generated and the conscious percept is adjusted accordingly. We can see exactly what happens when there is no such motor command, and hence no such adjustment, when we press on our eye with our finger. When we do this, the world itself seems to tilt and shake.

It took more than a hundred years for Helmholtz’s ideas to be applied to psychosis (Feinberg, 1978). Although Feinberg’s initial paper was on thought (which he took to involve ‘motor mechanisms’) and thought insertion, the easiest symptoms for

which to introduce the account are delusions of control, since it is clear that, if anything involves motor commands, bodily actions do.<sup>1</sup> In delusions of control, a subject may perform actions that are in keeping with her plans and intentions (for example, she might brush her hair), but she claims that somebody else is controlling her. Frith and Done (1989) took this to be a problem with self-monitoring. In particular, there is a mismatch between the predicted and actual sensory consequences of the bodily movement and so (as with Helmholtz's ocular example) the movement is attributed to an external source. This, in principle, could be a problem with the generation of the prediction itself (e.g. the corollary discharge) or with the mechanism that compares expected and actual sensory (including proprioceptive) input, what became known as "the comparator". Later manifestations of the self-monitoring theory (Frith, Blakemore, & Wolpert, 2000) saw comparator-based self-monitoring as the human body's way of meeting a computational challenge, in particular, involving skilled reaching and online correction (see Wolpert, 1997 for a review of these computational approaches to motor control). This connection, in the late nineties, between the computational neuroscience of healthy cognition and the neuropsychology of schizophrenia undoubtedly contributed to the credibility of SMTs.

### 1.2. Support for SMTs

Whereas in Helmholtz's example, the recognition by the nervous system that a certain stimulus is self-produced causes a correction of the conscious percept, in more typical bodily motor control, it results in sensory attenuation. The evolutionary benefit of this is clear enough: your nervous system needs to pay attention to stimuli that come from the outside, not the endogenous stimuli that (in a well-functioning system) will be harmless and irrelevant. Various data suggest that something goes wrong with this monitoring and subsequent attenuation (Blakemore, Smith, Steel, Johnstone & Frith, 2000). The most striking such datum is the reported finding that subjects with diagnoses of schizophrenia can tickle themselves. The postulated explanation for this is that there is a mismatch between expected and actual sensory consequences and the sensory consequences are not attenuated: the tickling sensation is like being tickled by somebody else. Typical subjects can't tickle themselves because their nervous systems accurately monitor, and successfully attenuate, the sensory consequences of the tickling movements (Blakemore, Wolpert & Frith, 1999).

It is not only with bodily action that support has been shown for the claim that self-monitoring goes awry in schizophrenia. Studies showed that patients with diagnoses of schizophrenia are more likely to misattribute their own voices than healthy controls (Cahill, 1996; Johns et al., 2001). For example, Johns et al. (2001) got subjects to read words aloud and played them feedback of their voices with mild acoustic distortions. The subjects with diagnoses of schizophrenia were considerably more likely than healthy controls to claim that the voice they heard was someone else's. Indeed, support that this was directly related to the production of hallucinations was supported by the result that, within the "schizophrenia" group, those who experienced hallucinations were more likely to misattribute their voice than those who did not hear voices.

### 1.3. Applying SMTs to AVH

Several theorists (Feinberg, 1978; Frith, 1992; Jones & Fernyhough, 2007; Seal et al., 2004) have attempted to explain AVHs in terms of inner speech misattribution, based on self-monitoring abnormalities. Although from a pre-theoretical standpoint it is not obvious that inner speech involves motoric elements, this has been empirically supported by several electromyographical (EMG) studies (which measured muscular activity during inner speech) some of which date as far back as the early 30s (e.g. Jacobsen, 1931). Later experiments made the connection between inner speech and AVH, showing that similar muscular activation is involved in healthy inner speech and AVH (Gould, 1948; McGuigan, 1966). The involvement of motoric elements in both inner speech and in AVH is further supported by findings from Gould (1950), who showed that when his subjects hallucinated, subvocalisations occurred which could be picked up with a throat microphone. That these subvocalisations were causally responsible for the inner speech allegedly implicated in AVHs, and not just echoing it (as has been hypothesised to happen in some cases of verbal comprehension (cf. e.g. Watkins, Strafella, & Paus, 2003)), was suggested by Bick and Kinsbourne (1987), who demonstrated that if people experiencing hallucinations opened their mouths wide, stopping vocalisations, then the majority of AVHs stopped.

To recap, SMTs have been presented as a unifying model for understanding the positive symptoms of schizophrenia. The idea is that there is one deficit concerning the monitoring of self-produced stimuli, and the symptoms differ because there are different kinds of self-produced stimuli. In delusions of control it is physical action, whereas in AVH and thought insertion it is inner speech.

## 2. Two challenges facing the application of SMTs to AVH

As mentioned, criticisms of SMTs tend not to take issue with the role of self-monitoring deficits in delusions of control, but rather concern the aforementioned extension of self-monitoring to symptoms that don't involve bodily action, namely, AVH and thought insertion. In this paper, I specifically focus on two challenges facing the application of SMTs to AVHs.

<sup>1</sup> Later I present a framework that does away with motor commands as traditionally construed, even for overt bodily actions.

### 2.1. The Auditory Phenomenology Challenge

How are we to account for the distinctly *auditory* phenomenology of certain AVHs? As [Cho and Wu \(2013\)](#) put it, if a theory claims that AVHs are misattributed inner speech, it must explain how we get a transformation from

the experience of the subject's own inner voice [...] often lacking acoustical properties such as pitch, timbre, and intensity into the experience of someone else's voice with acoustical properties. (p.2)

As [Wu \(2012\)](#) puts it in an earlier paper, “we must explain this ‘transformation’ from the normal to the pathological” (p.94). Although one can question the premise that inner speech lacks auditory phenomenology ([McCarthy-Jones and Fernyhough, 2011](#); [Moseley & Wilkinson, 2014](#)), even granting that this is the case, the PPF makes this ‘transformation’ less perplexing.<sup>2</sup>

### 2.2. The Varieties of AVHs Challenge

[Jones \(2010\)](#) emphasises the heterogeneity of AVHs:

The term AVH encapsulates a diverse phenomenological experience, which may involve single and/or multiple voices, who may be known and/or unknown, speaking sequentially and/or simultaneously, in the first, second, and/or third person and which may give commands, comments, insults, or encouragement. (2010, p.566)

In particular, Jones presents two models for understanding AVHs and argues that both are promising for understanding different subtypes of AVH.

The first model he examines is the “memory-based” model proposed by [Waters, Badcock, Michie, and Maybery \(2006\)](#), which he suggests fits well with the phenomenological features of AVHs that appear to be the upshot of a traumatic experience. In these AVHs you often find, for example, instructions to self-harm in a voice that is recognisably that of the abuser. Even more strikingly, the contents of these AVHs “can be related to what was said during or surrounding these events (e.g. if you tell anyone I’ll kill you)” ([Jones, 2010, p.568](#)). However, these form only a relatively small subset of AVHs (e.g. 4 out of 24 in [Fowler et al., 2006](#)). A larger subset of AVHs seem to serve the function of regulating or commenting on current events (e.g. [Nayani and David \(1996\)](#) reported that for 46% of their sample, the voice had replaced their “voice of conscience”). For these, [Jones \(2010\)](#) suggests that his second candidate model is a better option. This model is precisely the inner-speech-based SMTs that we introduced in Section 1.

The moral of this is that it is very difficult to account for all this variety within a model that explains AVHs in terms of misattributed inner speech. I am in complete agreement with Jones that we need different models for different subtypes.<sup>3</sup> I would simply add to Jones’s suggestion, however, that we distinguish frameworks, theories, and models.

Roughly, the distinction is as follows. *Frameworks* are very broad; they are ways of approaching a particular domain of inquiry (e.g. the brain and cognition). It is *within* them, whether implicitly or explicitly, that theories are built.<sup>4</sup> *Theories* are falsifiable claims (rather than ways of approaching something) at a high level of generality that explain a phenomenon or class of phenomena by elucidating the fundamental nature of the phenomenon, or by postulating certain rules or principles that govern or define the phenomenon (e.g. a self-monitoring theory of schizophrenia). *Models* are at a lower level of generality; they explain how a particular kind of phenomenon arises, sometimes in terms of other, hopefully better understood, phenomena (e.g. a misattributed inner speech model of AVH). Theories might suggest candidate models. So a self-monitoring theory of the primary symptoms of schizophrenia might suggest a misattributed inner speech model for AVHs occurring in the context of schizophrenia. Models should be applicable to individual cases. Thus, given a theory that posits one kind of deficit, different models could well be needed for different symptomatic manifestations of that underlying deficit.

In a nutshell, I will suggest that the different subtypes need (as [Jones, 2010](#) suggests) to be explained in terms of different aetiological models. However, these models are to be built within the same framework, the predictive processing framework that I will present now.

## 3. The predictive processing framework (PPF)

I’d like to put psychosis to one side, and focus, in this section, on presenting a framework for understanding the brain and cognition *generally*. We will later see how this might apply to AVH. I will start with an initial presentation of the PPF, and then present some conscious effects that are suggestive that the brain is engaged in predictive processing.

<sup>2</sup> One can also question the claim that AVHs always have auditory phenomenology. For example, [Frith \(1992, p.73\)](#) maintains that AVHs can involve “an experience of receiving a communication without any sensory component”.

<sup>3</sup> As Jones puts it, “[A]lthough models are created to fit the phenomenology of AVHs, due to the diverse nature of the experience they inevitably end up only fitting specific types of AVHs. Whilst it is possible a grand unified theory of AVHs and AHs may be developed, possibly involving a neurological mechanism cutting across many cognitive, metacognitive, linguistic and/or perceptual capacities, at present no such model exists” ([Jones, 2010, p. 573](#)).

<sup>4</sup> Although, of course, the adoption of frameworks will tend to be motivated by theoretical commitments. For example, the PPF is motivated by theoretical commitments concerning brain function. But this is a framework because it provides a way of approaching to a wide variety of issues.

### 3.1. An initial presentation: “inference” and efficiency

According to the PPF, the brain's main task is to “infer” from incoming signals, what the causes of those signals are, in other words, settle on a hypothesis about what is “going on”.<sup>5</sup> However, the incoming signal, namely, the proximal stimulus on sensory receptors, underdetermines distal causes. Since inputs are noisy and ambiguous, there is no one-to-one mapping: the same stimulation can be brought about by two very different distal causes (and different stimulation in different circumstances can be caused by the same distal cause). Given that more than one hypothesis is compatible with the incoming signal, how does the brain settle on one hypothesis rather than another? It needs to take two things into account: first, the fit of the input with the hypothesis, and, second, how statistically likely that hypothesis is (the “prior probability”), at least as far as the brain is concerned (this is *subjective* rather than objective probability). A hypothesis could fit the input extremely well, but its prior probability could be so low that it isn't even considered. Conversely, a hypothesis could have such a high prior probability, that, even though it doesn't fit the input well, it is settled upon.

Although this is a picture of what the brain has to do in order to do a good job of “inferring” what is going on, it is abstracted from the temporal dynamics of the brain's functioning. In reality, the inputs, hypotheses and prior probabilities (priors) are in constant hierarchical interaction with each other. This is where the notion of prediction comes in. What the selection of a hypothesis does is that it determines a set of predictions about subsequent inputs, namely, inputs that are compatible with the hypothesis. If the hypothesis does a good job of predicting inputs, it will be kept. If it does a bad job, it will be tweaked or abandoned altogether in favour of another hypothesis. In other words, one hypothesis is selected rather than another if it better *minimises prediction error*.

Some theorists (e.g. [Hohwy, 2013](#)) sum up the PPF by saying, for example, that the brain is in the business of minimising prediction error.<sup>6</sup> This prediction error minimisation is not only taken to account for perception and cognition, but for action as well (see e.g. [Adams et al., 2013](#)). Instead of there being motor commands, as on the standard picture, what you have are predictions, which are then fulfilled by the subsequent bodily movement, thereby also being a case of prediction error minimisation. This is often called “active inference” ([Friston, 2009](#)), which [Pickering and Clark \(2014\)](#) helpfully gloss as follows: “the combined mechanism by which perceptual and motor systems conspire to reduce prediction error using the twin strategies of altering predictions to fit the world and altering the world (including the body) to fit the predictions” (p.1).

This picture has interesting consequences for how we are to view the role of input on sensory receptors and its impact on higher cortical regions. According to the PPF, the only information that gets passed on up the hierarchy is *prediction error*. Indeed, as some recent theorists nicely put it:

an expected event does not need to be explicitly represented or communicated to higher cortical areas which have processed all of its relevant features prior to its occurrence.

[[Bubic, von Cramon, and Schubotz \(2010, p.10\)](#)]

This stands in sharp contrast to the standard (and admittedly intuitive) view of perception (what might be called a “bottom-up feature-detection view”). On such a view, inputs come in, are processed, and passed on. On the PPF, only the prediction error gets passed on, and therefore the vast majority of what determines your perceptual experience is what your brain has already predicted, your brain's best hypothesis. In short, on the PPF, the role of the incoming signal in determining conscious perception is much less great than on standard bottom-up views.

### 3.2. Precision weighting, second-order prediction and “attention”

Although incoming signals are ambiguous, in different contexts, the degree of ambiguity will differ. To maximise its predictions, the brain needs to accurately estimate how much ambiguity (uncertainty) there will be. In contexts where low ambiguity is expected, higher precision will be demanded, and *vice versa*. This is called “precision weighting” in the literature ([Feldman & Friston, 2010](#); [Friston, 2009](#); [Hohwy, 2012](#)), and is taken to be modulated by neurotransmitters such as dopamine ([Corlett, Taylor, Wang, Fletcher, & Krystal, 2010](#)).<sup>7</sup>

Recently, theorists ([Feldman & Friston, 2010](#); [Hohwy, 2012](#)) have equated this precision weighting with what we often call “attention”. As is well known, attention can be brought about endogenously or exogenously, namely, either because the subject is in a state of wanting to attend to something (endogenous attention), or because something in the environment attracts attention (exogenous attention).

Suppose I'm talking to someone in a crowded bar. By using information from seeing the person's lips move, my brain has certain expectations about what sounds are likely to be produced. That part of my visual field, and the corresponding part of

<sup>5</sup> One might say what is “out there”, but a great deal of information is internal to the body (see [Seth \(2013\)](#) on interoceptive predictive processing).

<sup>6</sup> To quote Hohwy's rather stronger claim (and one that I would not assent to without serious qualification): “prediction error minimisation is all the brain ever does” (2013, p.7).

<sup>7</sup> Theorists tend to take the mechanism of precision-weighting to involve specifically the prediction error rather than the predictions themselves. That is to say, the prediction error is given more weight in high-precision contexts. This amounts to the system being more demanding in terms of precision. However, the same effect could be achieved, not at the level of prediction error, but at the level of the precision of the priors (i.e. the predictions themselves). If the priors are more precise, then you are likely to get more prediction error. Distinguishing between these will not amount to a computational difference. It will (depending on how robustly realist one is about the PPF) either be an implementational difference, or simply to two ways of talking about the same thing. Either way, in contexts of high-precision, you will get either more prediction error, or prediction error that carries more “weight”.



my “auditory field”, now has relatively high precision attributed to it, largely because I am interested in what’s going to be said (endogenous focal attention). Everything else around that, both visual and auditory, has relatively low precision. My brain has a very gist-like hypothesis about what is going on there, and, any prediction error regarding that has low weighting. Of course, a loud peripheral bang or camera flash can grab my attention (a case of exogenous focal attention). That would constitute a strong enough input to overcome whatever down-modulation my brain has placed on that peripheral prediction error.

To sum up, then, we can think about the directing of attention as “turning up” the expected precision of an incoming signal, which amounts to turning up the “gain” on any prediction error. This can be viewed normatively. That is to say, not only can your brain get its predictions wrong, it can also get the expected precision of its predictions wrong, i.e. its second-order predictions. Your brain can wrongly think that it is in a more or less ambiguous environment than it actually is in. And this will mean too much or too little weighting on prediction error.

### 3.3. *Conscious effects of predictive processing*

#### 3.3.1. *Binocular rivalry*

This is perhaps the most common example used in support of the PPF. It speaks strongly against a bottom-up view in favour of a more predictive or “inferential” view. Under experimental conditions, each eye is presented with a different, but meaningful, stimulus. One standard example involves presenting one eye with a picture of a house, and the other with a picture of a face. Subjects do not report visually experiencing, as one might expect, a mixture of face and house. Rather, they experience a “bi-stable” switching, from face to house, and back, and so on (the switching is often reported as a gradual “breaking through” of the other image). As [Hohwy, Roepstorff, and Friston \(2008\)](#) point out, this can be nicely explained within the PPF, as a reasonable response to highly un-ecological circumstances. In a nutshell, you experience the bi-stable state because your brain is switching between hypotheses about what is out there. When it settles on one hypothesis, say, the face hypothesis, inputs from the house image fail to accord with this hypothesis and prediction error is sent up the hierarchy. When enough prediction error accumulates, the hypothesis switches (and with that, what the subject consciously experiences) to the house hypothesis, but then the input from the face image doesn’t accord, and so on, and so forth. The fact that this bi-stable switching only occurs with certain well-chosen stimuli, suggests the operation of what might be called a “hyperprior”: an expectation about the world that is stable, and often at a high degree of abstraction. The “hyperprior” in this case is that faces and houses, being at different scales, cannot occupy the same region of the visual field at the same time. As a result, the “face-house” hypothesis about what is out there never presents itself.

Note how binocular rivalry puts pressure on a “bottom-up” view of perceptual experience. As [Hohwy \(2013\)](#) puts it:

During rivalry, the physical stimulus in the world stays the same and yet perception alternates, so the stimulus itself cannot be what drives perception (p.20)

One might think that binocular rivalry could result from the operation of a different principle, and one that is compatible with a largely “bottom-up” view, namely the principle: “If your eyes each have coherent but incompatible inputs, don’t receive input from both at once.” This nicely illustrates how predictive processing concretely differs from a bottom-up view such as this one. When the subject experiences a house, on the bottom-up eye-selection hypothesis, this means that the eye being presented a house is passing on a signal, whilst the eye being presented a face is not. On the predictive processing view, it is the opposite: the eye being presented the face is passing on a signal, namely a prediction error signal.

One reason for rejecting the eye-selection account is that it doesn’t (unlike the build-up of prediction error) explain the stable switching back-and-forth. A second, more striking, reason is that this account was falsified by [Diaz-Caneja \(1928\)](#) who discovered that if each stimulus picture is cut in half and swapped over, so that each eye is being presented with a half-face half-house image (split down the middle) there is the same experience as before, namely, a rivalry between a complete percept of a face and a house.

#### 3.3.2. *The Hollow Mask Illusion*

When you are presented with a rotating mask that is slowly turned to present you with the concave back of the mask, your brain “corrects” the concave stimulus into a convex stimulus. (You yourselves can experience this illusion on several sites on the internet.) You experience the concave back of the mask, as convex. Again, this is due to a very strong prior (perhaps worthy of being called a hyperprior), that over-rides the incoming signal. The prior in question is that the faces you will encounter will always be convex (a fair expectation!). This prior is so strong that even though the “concave face” hypothesis would better match the input, it is never selected. Given the highly specialised processing that faces receive in human cognition, this is perhaps hardly surprising.

#### 3.3.3. *The McGurk Effect*

Moving away now from uni-modal priors, there is an effect that deals with highly flexible, concrete, cross-modal priors. This is the fact that what you see will affect your brain’s expectations about what it hears. If a subject is played a video of someone appearing to say “ga, ga”, but you play them a synchronous audio track of someone saying “ba, ba”, you aurally experience “da, da” (something “in between”). This is called the McGurk Effect. The visual input changes your brain’s expectation concerning the auditory input, and interprets it differently.

We can see the same cross-modal predictive processing (vision having an effect on audition) at work in more everyday cases. If you look across a crowded bar at someone ordering a drink, by looking at her lips, you can actually *hear* her order. Your brain uses the visual input to inform its priors about what is being said. If you had tried to hear her order with your eyes closed, there is no way you would have been able to single out her voice from the noise of the crowded room. Your brain just wouldn't have had the visual cues to inform its expectations.<sup>8</sup> Furthermore, since what she says is what matters to you at that moment, you “turn up the gain” on prediction error that is relevant to hypotheses concerning what she says.

### 3.4. Clearing up potential confusion: what the brain does and what the person does

Although this predictive processing is extremely relevant for our conscious experience, we need to be very careful to distinguish what our brains do from what we do. One nice way of seeing this distinction is by distinguishing “surprisal” (Tribus, 1961) and “surprise”, where the former is what is surprising to your brain (as signalled by prediction error), and the latter is what is surprising to you.<sup>9</sup> These two come apart. If you look out of your window and see an elephant on the lawn, you might be very surprised. However, the fact that you see straight away that it is an elephant shows that your brain has already minimised the prediction error and settled on the elephant-on-lawn hypothesis. Conversely, when you are attempting a binocular rivalry task, and you've done it before, the switching doesn't surprise you, however, your brain is switching between hypotheses precisely because it is struggling to keep “surprisal” to a sufficiently low level. Since we are concerned with AVHs, a conscious phenomenon, it is vital to understand this relationship between what the brain does, and what the person does, and, analogously, predictive processing on the one hand, and conscious experience on the other. Roughly, your conscious percept is determined by the overall hypothesis that your brain has adopted in order to minimise prediction error.

## 4. Predictive processing and psychosis

Now that I have presented the PPF, we can see what relevance it has for psychosis in general, and for AVH in particular. Two sources of evidence that support the hypothesis that predictive processing abnormalities might be implicated in psychosis are, first, that the conscious effects that have been taken to be suggestive of predictive processing are altered in subjects with diagnoses of schizophrenia, and, second, that there are behavioural results from eye tracking that are nicely explicable in terms of predictive processing.

### 4.1. Conscious effects are altered in patients with diagnoses of schizophrenia

Suppose that psychosis can be understood, in part, at least, as a breakdown in normal predictive processing. And suppose that, as we have suggested, the conscious effects just mentioned are the product of normal predictive processing. One would expect patients with schizophrenia diagnoses to experience these effects differently from the normal population. This is exactly what the evidence suggests.

Schneider, Leweke, Sternemann, Weber, and Emrich (1996), Schneider et al. (2002) and Emrich, Leweke, and Schneider (1997) demonstrate that subjects with schizophrenia diagnoses do not experience the Hollow Mask Illusion. They see the mask, correctly, as hollow. Their brain does not “correct” the input.

Pearl et al. (2009) showed that patients with diagnoses of schizophrenia experience the McGurk Effect less. In other words, the visual stimulus corrects the auditory input less, or not at all. Pearl et al. (2009) put this in terms of a “decreased reliance” on visual cues. Results with similar implications are reported by Ross et al. (2007) with regards to visual enhancement of speech comprehension in noisy environmental conditions. In other words, subjects with diagnoses of schizophrenia showed an impairment in their ability to enhance their auditory perception with visual cues. Within a PPF this would be put in terms of either decreased efficacy of auditory priors informed by the visual information, or in terms of excessively weighted bottom-up prediction error (or both).

As for binocular rivalry, Heslop (2012) showed that switching rates in binocular rivalry were, on average, significantly slower in subjects with schizophrenia (0.28 switches per second, versus 0.54 in the non-clinical population).

### 4.2. Evidence from eye tracking

Following Adams, Perrinet, and Friston (2012), I distinguish three eye-tracking tasks where the performance of patients with diagnoses of schizophrenia differs significantly from healthy controls.

First, there are cases of impaired tracking during visual occlusion (Hong, Avila, & Thaker, 2005; Thaker et al., 1998). In other words, when a moving target is temporarily occluded from view, controls are much better than subjects with diagnoses of schizophrenia at keeping track of the target when it comes back into view.

<sup>8</sup> Similarly, playing loud audio recordings of crowded rooms was used as a simple counter-surveillance technique. Two people can easily converse by looking at each-other's mouths producing the sounds. Their brain uses predictions generated by the visual information to pick out the relevant auditory information from the noisy background. Anyone listening in on a bug will lack the necessary visual cues to pick out the conversation from the noise in the same way.

<sup>9</sup> There's obviously a similar distinction to be drawn between your expectations (say, when placing a bet) and your brain's priors.



Second, patients with diagnoses of schizophrenia show impaired “repetition learning.” When a target trajectory is repeated, healthy controls achieve optimal performance whereas subjects with schizophrenia diagnoses do not (Avila, Hong, Moates, Turano, & Thaker, 2006).

Thirdly, there are cases of what, in the literature, is called “paradoxical improvement”, namely, where patients with a supposed illness or deficit perform better at a particular task than healthy controls (the fact that patients with schizophrenia do not experience the Hollow Mask Illusion might be thought of as a “paradoxical improvement” since their percept is more veridical). The task in question involves keeping track of a target that rapidly and unexpectedly changes direction. Subjects with schizophrenia are better at keeping track of the target than healthy controls (but only very shortly after the change in direction).

Working within the PPF, Adams et al. (2012) explain these three discrepancies in terms of differing reliance on predictions (priors) and prediction error respectively. The first two tasks are improved by reliance on prediction, whereas the third task, being deliberately unpredictable, is hindered by prediction, and would be improved by a higher weighting on the prediction error. Note, however, that, like the Hollow Mask Illusion, the “paradoxical” improvement is shown on a stimulus that is statistically un-ecological. Just as we encounter convex faces and never concave ones, the world is full of statistical regularities and it is more adaptive to be able to exploit those regularities efficiently than to slightly outperform others in the rare instances when stimuli are totally unpredictable.

## 5. Addressing the two challenges

I will start by showing how the PPF addresses the *Auditory Phenomenology Challenge*. In particular, I will show how it circumvents the issue of explaining the “transformation” from the healthy phenomenon (e.g. inner speech) to the pathological phenomenon (i.e. AVH). I will then present three subtypes of AVH and show how the PPF accommodates different aetiological models for each.

### 5.1. Dissolving the Auditory Phenomenology Challenge

Viewed from within the PPF, this challenge is based on a mistaken understanding of how the brain works, and the relationship that this has to conscious experience.

The PPF changes how one thinks of perceptual experience, and, by extension, radically changes one’s explanatory focus in trying to account for hallucinations. On a standard framework, where front-line sensory stimuli get gradually processed and passed on up the hierarchy, hallucinations make one wonder, “Where does this erroneous sensory stimulus come from?” Indeed, we can see SMTs as making attempts to answer this within a standard framework. Their answer is: they come from the quasi-sensory stimulation of inner speech, which is then misattributed. However, when, instead, we adopt the PPF, incoming stimuli play a much smaller role in determining the conscious percept, even where veridical perception is concerned. Given that a conscious percept is constituted by the hypothesis that best minimises prediction error, we don’t ask, “Where does the input come from?”, since the input alone doesn’t (and can’t) determine the percept. Rather we ask, “Why does this hypothesis minimise prediction error?” This general approach makes hallucinations both less perplexing, and less different from veridical perception.<sup>10</sup> The point is that on the PPF, your conscious experience at any given time is the hypothesis that your brain has selected in order to minimise prediction error. Applied to AVHs, your experience will have auditory phenomenology if your brain has had to adopt the hypothesis that you are hearing something in order to minimise prediction error.

This isn’t really a “transformation” because, on the PPF, it makes no sense to talk of inner speech as a “raw material” that needs transforming. Either you have an experience that is inner speech, because your brain has adopted the hypothesis that that is what is going on, or you have a different experience, corresponding to a different hypothesis. There is no *experience* of inner speech *first*, which is somehow then transformed. The question about whether inner speech is relevant for AVHs is not one of transformation. It is about whether relevant elements involved in the production of the inner speech experience are also involved in the production of some AVHs.<sup>11</sup> It seems fairly clear, given all of the evidence in its support (some of which we have already seen, some of which we will touch upon below), that the answer to that is: yes.

However, it is worth noting that on the PPF, all action will be couched in terms of active inference, namely, in terms of (hopefully) self-fulfilling sensory predictions (cf. Adams et al., 2013). How we are to think of inner speech within this framework is a fascinating direction for future inquiry. Given the plausible view that inner speech is developmentally derived from overt, private, speech (Berk, 1992; Vygotsky, 1934; Winsler, 2004), this could involve activating a prediction (as for any overt action), but over time somehow managing to dispense with many of the features of the action in question. This could involve

<sup>10</sup> In some respects, Helmholtz was quite right when he said that “Perception is controlled Hallucination”. This is trivially false if “hallucination” is understood normatively, as being *by definition* inaccurate: Helmholtz meant this to apply to fully veridical perceptual experience. What he meant was that, to put it bluntly, the vast majority of the “raw materials” for our perceptual experiences are prediction that originate in our heads, with an important, but informationally tiny role, given to incoming signals.

<sup>11</sup> This resembles what Wu (2012) calls “the moderate identification thesis”, which he claims to not be taking issue with. The strong identification thesis, which he criticises on empirical grounds, I would criticise on conceptual grounds. It’s not *even* false: it’s confused. You can’t say that AVHs just are episodes of inner speech. One experience cannot be the same as a different experience!

learning to lower the weighting on the prediction error, or to make the prediction itself less demanding, or a combination of the two. Either way, prediction error could be minimised without having to go through with all aspects of the action (or, rather, its developmental precursor, private speech), although some aspects (e.g. subvocalisations, muscular activations) will remain. How this could lead to AVH is something that we will come to shortly. The key idea, roughly, is that inner speech is stripped down outer speech, but that this is built back up in AVH. However, it is not simply built back to overt speech, but to something different that involves the hearing of speech, but not its production.

## 5.2. Accounting for three subtypes of AVH

Jones (2010) expressed a concern that one model is unlikely to explain the various subtypes of AVH. He presented two subtypes. I will add a third but then gesture towards how the PPF provides a framework against which all three can be explained through different aetiological models.

### 5.2.1. Inner speech hallucinations

As we mentioned, the most popular model for understanding AVHs involves misattributed inner speech, where the misattribution results from an abnormality in self-monitoring. Although I want to abandon the framework within which SMTs tend to function, following Jones (2010), I think it is important to emphasise that there is great deal of empirical support for the view that a subset of AVHs implicate mechanisms involved in inner speech.

There is support for the idea that inner speech might be relevant, other than the experiments we noted earlier in support of SMTs. Ford and Mathalon (2004) used EEG findings to suggest that during the production of healthy inner speech, corollary signals from speech production areas seem to suppress activation in speech processing areas. As Bentall and Varese (2013) nicely put it, “it might be said that when we talk to ourselves, the frontal language production areas of the brain tell the posterior areas not to bother listening” (2013, p.71). This attenuation should remind us of the support from bodily effects of monitoring, namely, the attenuation of sensation in (attempted) self-tickling. Both the attenuation pertaining to self-tickling and to inner speech can be accounted for within the PPF.

Within the PPF, the fact that we cannot tickle ourselves is explained rather differently from how it is explained in SMTs that rely on motor commands and self-monitoring mechanisms. Brown and et al. (2013) have done work within the PPF that accounts for this in terms of a more general dampening of weighting on prediction error. The reason for this dampening is that movement involves implementing a currently false hypothesis (“I am moving”) in favour of the currently true hypothesis (“I am not moving”). The dampening swings the balance in favour of the false hypothesis, which then becomes true when the action is initiated.<sup>12</sup> In inner speech, there is the prediction that you will speak, but the precision weighting will be turned right down, circumventing the need to minimise prediction error by actually overtly speaking (although, as we saw, some vestiges of the overt speech will remain). This will correspond to the suppression of activation in “speech processing areas” which are reported in both inner speech and in hearing oneself speak aloud (Ford & Mathalon, 2004). On the PPF there would be less activation in these areas, even when hearing oneself speak aloud, since there would be less prediction error (better predictions) than hearing someone else speak.

The fundamental principle of prediction-error minimisation is the same whether the activity to be predicted is self-generated or not. One important difference, however, is that, if the activity is self-generated, the expected precision will be higher. Given this, only a relatively slight error in prediction would result in highly weighted prediction error, which would subsequently result in a more erroneous hypothesis to explain it away. This could account for cases where there are delusions of control and inner speech-based hallucinations in the absence of more outwardly directed perceptual abnormalities.

### 5.2.2. Memory-based hallucinations

Waters et al. (2006) distinguish their view from “standard cognitive accounts”, by which they mean SMTs, and which they call “single deficit accounts”, by proposing that:

at least two cognitive deficits must be present to explain auditory hallucinations: (1) a fundamental deficit in intentional inhibition which leads to auditory mental representations intruding into consciousness in a manner that is beyond the control of the sufferer; and (2) a deficit in binding contextual cues, resulting in an inability to form a complete representation of the origins of mental events.

[Waters et al. (2006, p.66)]

This talk of “auditory mental representations”, “inhibition” and “binding of contextual cues” is nicely illustrative of how removed this is from the sort of framework that I am suggesting in this paper. However, this is not to say that they are incompatible, and cannot both be helpful.

Using a couple of tasks (the Hayling Sentence Completion Task (Burgess & Shallice, 1996) and the Inhibition of Currently Irrelevant Memories Task (Schneider & Ptak, 1999)) Badcock, Waters, Maybery, and Michie (2005) demonstrated that hallucinating patients with diagnoses of schizophrenia have lower inhibition (more intrusion) of irrelevant associations and

<sup>12</sup> So both PPF and SMTs have an account of why we cannot tickle ourselves, but is there anything to adjudicate between them? Well, they do generate different experimental predictions. SMTs view ticklishness as a function of unpredictability of specific movements that would be rendered predictable through self-monitoring. PPF on the other hand, views it as a more general dampening of precision on prediction error.

memories. The second deficit is needed to explain why these memories, which are failing to be inhibited, aren't being recognised as memories. In research on episodic memory, there is a distinction between content (what is remembered) and context (e.g. temporal and “source” features of what is remembered). Waters et al. (2006) suggest that, although the content is remembered, there is a deficit in “context memory”, and, more specifically, “an impairment in combining contextual cues together to form an integrated representation of an event in memory”. Thus, the memory will present something that happened in the past, usually something traumatic, but it will not be experienced as a memory. It will be taken as a voice from the present.

Let us look at what might be going on here through the lens of the PPF. In a case of an episode of healthy episodic memory, whether deliberate or unbidden, what is involved is the activation of relevant imagery, which within the PPF (in a way somewhat similar to our earlier discussion of inner speech) is not some kind of inner sensation, but a prediction. However, unlike the predictions at play in perception and action, this prediction is not answerable to – malleable in the light of – inputs. How is this “decoupling” from the environment achieved? What needs to happen is that the prediction need not to be “taken too seriously”, and so this will mean turning the precision right down on prediction error (this is similar to what happens in inner speech, minus the articulatory component). That way the “prediction” can be maintained even though the system is fully aware that it is not really predicting anything, that there is nothing really corresponding to the localised hypothesis generating the prediction. The general, overarching, hypothesis, is still that the subject is where she is (e.g. at a desk), but she is simply remembering something (e.g. that time when her father shouted at her).

Now suppose that this prediction is activated, but there is a problem with keeping the weighting on prediction error low. This generates erroneous amounts of prediction error for which the brain has to adopt a hypothesis: this is perception, not memory. In a sense, then, if predictive processing, and, in particular the upwards and downwards modulation of precision on prediction error, goes badly awry then we get a blurring between mere imagery (as involved in episodic memory, imagination and, with relevant qualification, inner speech) and perception. The clever trick of decoupling, which enables us to remember episodically, or to imagine vividly, is disrupted.

However, if this is correct, then we don't need to appeal to integration of contextual cues (except insofar as there is a *de facto* failure to realise that this is not happening now). The recollective episode is not recognised by the subject as a memory, not because of some cognitive failure to remember the temporal or “source” context in which this happened, but because it doesn't feel phenomenologically like remembering something: it feels like hearing something. Indeed, failure to remember temporal or source context cannot suffice to explain AVHs since there are cases where the voices are recognised by the subject as being exact replays of the past (12% in McCarthy-Jones et al., 2012), but they are still not experienced as *memories*, but as perceptual experiences.

Indeed Wu's *Auditory Phenomenology Challenge* applies to memory-based context-monitoring views almost as much as to inner-speech-based self-monitoring views. I say “almost as much” since perhaps (and we know very little about whether this is actually the case) episodic auditory memories have more phenomenological features in common with auditory perceptual experiences than episodes of inner speech (which arguably have an active, articulatory component). But, in any case, they certainly do not share all of them. So the challenge goes: How does one account for the auditory phenomenology of AVH if they are simply mis-identified memories? How does one account for the “transformation” from auditory-memory phenomenology to auditory-experience phenomenology? On the PPF account, since the conscious experience is a product of prediction-error minimisation, this question doesn't arise. The initial event, the explanatorily relevant genesis of the experience, may be (as with the inner speech hallucinations) the same initial event as for a recollection in episodic memory, however, the erroneous prediction error causes the brain to hypothesise that something very different is going on.

### 5.2.3. Hypervigilance Hallucinations

The two varieties I have just elaborated on were introduced earlier in the paper through a consideration of Jones's *Varieties Challenge*. This third kind has not yet been presented in this paper.

Perhaps the first step towards an appreciation of the existence of hypervigilance hallucinations was made in a paper by Delespaul, DeVries, and Van Os (2002), who conducted an investigation into the contextual influences on AVHs. They found that the context where voice hearing is most likely to occur is either in the presence lots of people or alone. Building on this, Dodgson and Gordon (2009) suggested, on grounds of clinical case-studies, a kind of AVH called a “hypervigilance hallucination.” Here is a description of hypervigilance hallucinations from their paper:

Michael was experiencing a series of stressors including break up with a girlfriend and exclusion from seeing their child, heavy street drug use, including amphetamine, and a pending court case for arson. He began to experience auditory hallucinations, hearing people calling him a “nonce”. Michael began to believe that people could read his thoughts, and that people thought he was a paedophile.

[Dodgson and Gordon (2009, p.328)]

The existence of hypervigilance hallucinations as a separate subtype was subsequently supported by Garwood, Dodgson, Bruce, and McCarthy-Jones (2013) who, based on a cluster analysis, showed that AVHs tend to occur when:

- (i) attention is directed *inward* in *quiet* contexts
- (ii) attention is directed *outward* in *noisy* contexts

They took (i) to suggest that inner speech hallucinations were occurring, and (ii) to suggest that hypervigilance hallucinations were occurring. Attention seems like a key factor here. So let's view this notion of hypervigilance from within the PPF and its gloss on attention as precision-weighting.

Let us contrast vigilance, with an illustratively opposing state: total calmness and lack of threat. Compare two cases of walking down a familiar woodland path at dusk. In the one case, you are walking home after an agreeable dinner party at a friend's house, and you are feeling happy and relaxed. Because it is getting dark, you will find that your brain is forming rather vague hypotheses about what is out there, but since you are relaxed, and you have no reason to think that anything might be a threat, this is fine: precision-weighting ("attention") is kept low. You get home safely having represented your environment in a vaguer, more coarse-grained way than you would have done in broad daylight, but this served your navigational purposes. Now suppose, instead, that you are walking home after having watched a horror movie. You are no longer relaxed, but in an emotional state of *vigilance*. As a result, you will be more likely to interpret something as looking like a person lurking amongst the trees. If this startles you, your attention will focus on that part of your visual field, and the precision-weighting of prediction error in that area will be turned right up. You may, due to this upwards-modulation, see that it is, in fact, only a tree trunk. Although as you look more closely, you get more incoming visual information suggesting the tree-trunk hypothesis, *interoceptive* information, namely, the fear that you feel, which is residual from the horror movie you saw, counts in favour of the lurking-person hypothesis (Pezzulo, 2013). You (or, perhaps more accurately, your nervous system) roughly think: "Why am I scared? There must be something to be scared of", and then you get even more scared. In sharp contrast, in the first case, when you are happy and relaxed, the lurking-person hypothesis doesn't even present itself.

This fits nicely with hypervigilance hallucinations. In a state of hypervigilance, ambiguous inputs will be given a threatening interpretation; hypothesis selection will be biased towards something threatening, because this will also serve the purpose of explaining the *interoceptive* state (Seth, 2013). For the subject, this will give rise to experiences which misrepresent reality: from the ticking clock or the muffled sound of neighbours talking will emerge the experience of a voice telling the subject exactly what the subject, in his state of hypervigilance, is afraid of hearing (e.g. "nonce!").

Unlike the two other subtypes, hypervigilance hallucination have their genesis in the external world. One might of course claim that, since hypervigilance hallucinations do, strictly speaking, have some kind of worldly stimulus, then they are not, *strictly speaking*, hallucinations; they are rather illusions. I see this as a mainly terminological point, however, so will not go into it.<sup>13</sup> A more important point is that, because of this, SMTs cannot account for hypervigilance hallucinations whereas the PPF can. The point is that, according to SMTs, normal monitoring exploits mechanisms (e.g. motor commands) linked to the production of self-generated events, and it is this that goes wrong and is misattributed. Thus, although it has a story to tell about how bodily actions, inner speech, and perhaps even episodic memories, can be badly monitored, and hence misattributed, it must remain silent about hypervigilance hallucinations, which are precisely not a case of a self-generated stimulus being misattributed, but rather an external stimulus being misinterpreted. To put it another way, unlike with SMTs, in the PPF, *all* stimuli, not just the self-generated ones, need to be predicted.

### 5.3. Multiple models within one framework: personal history and attentional focus

I have gestured towards how these three subtypes could be accounted for within the PPF, but why would a given subject, at a given time, experience one subtype rather than another? This is an unavoidably difficult question. Although I cannot answer it in full, I think that two things are of particular explanatory relevance within the PPF. The first is the personal history of the subject; the second is the subject's attentional focus, which will probably often be the accompaniment of certain moods, emotions and contexts.

Personal history, for example, whether a subject underwent some form of traumatic abuse as a child, is likely to have an impact on whether a subject hears voices, and, as we have suggested, these voices are perhaps likely to be memory-based. Similarly, if a subject (like Michael, in Dodgson and Gordon's case-study) has a history of amphetamine use, coupled with other so-called "stressors", this might well lead to heightened risk of AVH, but probably not memory-based ones, since there is no single traumatic event to form the basis of the relevant episodic memory. Of course there could be interaction between the different factors that contribute to the (perhaps unrealistically clear) delineations of the subtypes. Thus childhood trauma and substance abuse could contribute to someone being at heightened risk of AVH. These might be, for example, hypervigilance hallucinations whose content could in some way be tied to past trauma.

What, then, would make an individual more likely to experience AVHs that have an internal (viz. inner speech or memory) or an external (viz. hypervigilance) genesis? In answer to this question it may be useful to appeal to attentional focus. We saw that, within the PPF, following Hohwy (2012), we can think of attention as turning up the precision-weighting, and hence potential prediction error, on the attended stimulus. Attention can be directed outward, at incoming stimuli, or it can be directed inward, at thoughts, feelings and memories. Thus, if attention is directed outward in a state of anxious hypervigilance, there is more likely to be excessively weighted prediction error from the outside. This would correspond

<sup>13</sup> One can still be justified in calling them hallucinations. The distinction between hallucination and illusion is not necessarily one between two different loci of initial stimuli, but between two different experiences and how they represent relevant aspects of the world. In the one, something is represented which isn't there at all. In the other, something is presented but it is distorted. Viewed in this way, the experience in hypervigilance hallucinations is unrecognizable from the causal source of the stimulus. Something that is totally absent is being presented to the subject, regardless of whether something actual in the world is causally implicated. That cause is no longer in any way represented in the content of the experience.

to hypervigilance hallucinations. As a result, traffic noise, or a ticking clock, or the sound of a group of people talking, might be embellished into the experience of a voice saying something. Alternatively, if attention is directed inwards, and the subject is socially isolated, ruminating in a state of shame or guilt, then there is more likely to be excessively weighted prediction error from the inside. This would correspond to inner speech or memory-based hallucinations. This nicely fits the findings by Garwood et al. (2013), concerning the prevalence of hallucinations in quiet and noisy contexts. It also allows that one patient, in different contexts and different moods, may experience different AVH subtypes. Indeed all three of the subtypes presented here may occur in one subject at different times. Of course, the picture is more complicated, but the framework is precisely the sort that is capable of accommodating such complications on a case-by-case basis.

## 6. Summary and conclusion

I have suggested that we view AVHs through the lens of a predictive processing framework as a way of addressing two challenges, the first of which, in particular, plagues orthodox SMTs, the second of which is problematic for any single model of AVHs.

The first challenge, of how the auditory phenomenology of AVHs can be accounted for, or at least rendered less perplexing, is addressed by the fact that the subject's conscious percept is not determined by incoming stimuli, but by the brain's best hypothesis as selected on the basis of how well it minimises prediction error. The question is not: Where does this erroneous stimulus come from? Rather the question is: Why does the brain adopt such a strange hypothesis? The answer to this question is (at least partially): erroneous precision weighting on prediction error.

The second challenge, concerning the varieties of AVH, is addressed by realising that a new framework can accommodate elements of pre-existing models, but views the contribution of, for example, episodic memory or inner speech, in a way that is somewhat different. They are not "raw materials" to be transformed. Rather, mechanisms that are implicated in generating healthy experiences of episodic memory and inner speech, can generate erroneous prediction error that will result in erroneous hypotheses being adopted to minimise it. A major advancement of the PPF over SMTs is the realisation that, since prediction lies at the heart of cognition (broadly construed to include emotion and action), both internally and externally produced stimuli need to be predicted. SMTs had to be silent regarding external stimuli, since only self-generated stimuli can be monitored.

Adopting the PPF, like adopting any framework, is only a first step. Not only is there, first of all, work to be done on better understanding predictive processing; for example, its neurobiological underpinnings and how it might go wrong at that level (see Corlett et al., 2010 for advancements in this direction). There is also exciting work to be done within the PPF with specific case-studies. It is a flexible framework within which we can better understand unique individuals, with different histories, problems and conditions.<sup>14</sup>

## Acknowledgments

I would like to thank Charles Fernyhough, Pete Moseley, and Matthew Ratcliffe for helpful comments and discussions on earlier versions of this paper. This research was supported by a Wellcome Trust Strategic Award (WT098455MA).

## References

- Adams, R., Perrinet, L., & Friston, K. (2012). Smooth pursuit and visual occlusion: Active inference and oculomotor control in schizophrenia. *PLoS ONE*, 7(10).
- Adams, R., Shipp, S., & Friston, K. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, 218, 611–643.
- Avila, M., Hong, L., Moates, A., Turano, K., & Thaker, G. (2006). Role of anticipation in schizophrenia-related pursuit initiation deficits. *Journal of Neurophysiology*, 95, 593–601.
- Badcock, J. C., Waters, F. A. V., Maybery, M. T., & Michie, P. T. (2005). Auditory hallucinations: Failure to inhibit irrelevant memories. *Cognitive Neuropsychiatry*, 10(2), 125–136.
- Beavan, V., Read, J., & Cartwright, C. (2011). The prevalence of voice-hearers in the general population: A literature review. *Journal of Mental Health (Abingdon, England)*, 20(3), 281–292.
- Bentall, R., & Varese, F. (2013). Psychotic hallucinations. In F. Macpherson & D. Platchias (Eds.), *Hallucination: Philosophy and psychology*. Cambridge, MA, USA: MIT Press. ISBN 9780262019200.
- Berk, L. E. (1992). Children's private speech: An overview of theory and the status of research. In R. M. Diaz & L. E. Berk (Eds.), *Private speech: From social interaction to self-regulation* (pp. 17–53). Hove, UK: Lawrence Erlbaum Associates.
- Bick, P., & Kinsbourne, M. (1987). Auditory hallucinations and subvocalizations in schizophrenics. *American Journal of Psychiatry*, 14, 222–225.
- Blakemore, S.-J., Frith, C. D., & Wolpert, D. W. (1999). *Journal of Cognitive Neuroscience*, 11, 551–559.
- Blakemore, S. J., Smith, J., Steel, R., Johnstone, E. C., & Frith, C. D. (2000). *Psychological Medicine*, 30, 1131–1139.
- Brown, H. et al (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*.
- Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, 4(25), 1–15.
- Burgess, P. W., & Shallice, T. (1996). Bizarre responses, rule detection and frontal lobe lesions. *Cortex*, 32(2), 241–259.
- Cahill, C. (1996). Psychotic experiences induced in deluded patients using distorted auditory feedback. *Cognitive Neuropsychiatry*, 1(3), 201–211.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03), 181–204.
- Cho, R., & Wu, W. (2013). Mechanisms of auditory verbal hallucination in schizophrenia. *Frontiers in Schizophrenia*, 4.
- Corlett, P. R., Taylor, J. R., Wang, X. J., Fletcher, P. C., & Krystal, J. H. (2010). Toward a neurobiology of delusions. *Progress in Neurobiology*, 92(3), 345–369.

<sup>14</sup> I'd like to mention the possibility that the PPF might be useful for understanding voice hearing in other disorders (e.g. in major depression, bipolar disorder or PTSD) (Larøi et al., 2012), and, indeed, in the non-clinical population (Beavan, Read, & Cartwright, 2011; Larøi, 2012).



- Delespaul, P., DeVries, M., & Van Os, J. (2002). Determinants of occurrence and recovery from hallucinations in daily life. *Social Psychiatry and Psychiatric Epidemiology*, 37, 97–104.
- Diaz-Caneja, E. (1928). Sur l'alternance binoculaire [On binocular alternation]. *Annales d'Oculistique*, 165, 721–731.
- Dodgson, G., & Gordon, S. (2009). Avoiding false negatives: Are some auditory hallucinations an evolved design flaw? *Behavioural and Cognitive Psychotherapy*, 37(3), 325–334.
- Emrich, H., Leweke, F., & Schneider, U. (1997). Towards a cannabinoid hypothesis of schizophrenia: Cognitive impairments due to a dysregulation of the endogenous cannabinoid system. *Pharmacology Biochemistry and Behavior*, 56, 803–807.
- Feinberg, I. (1978). Efference copy and corollary discharge: Implications for thinking and its disorders. *Schizophrenia Bulletin*, 4, 636–640.
- Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 4.
- Fletcher, P. C., & Frith, C. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1), 48–58. <http://dx.doi.org/10.1038/nrn2536>.
- Fowler, D., Freeman, D., Steel, C., Hardy, A., Smith, B., Hackman, C., et al (2006). The catastrophic interaction hypothesis: How do stress, trauma, emotion and information processing abnormalities lead to psychosis? In W. Larkin & A. P. Morrison (Eds.), *Trauma and psychosis: New directions for theory and therapy* (pp. 101–124). London: Routledge.
- Ford, J., & Mathalon, D. (2004). Electrophysiological evidence of corollary discharge dysfunction in schizophrenia during talking and thinking. *Journal of Psychiatric Research*, 38(1), 37–46. Jan.
- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1456), 815–836.
- Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13(7), 293–301.
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews: Neuroscience*, 11(2), 127–138.
- Frith, C. (1992). *The cognitive neuropsychology of schizophrenia*. Psychology Press.
- Frith, C., Blakemore, S.-J., & Wolpert, D. M. (2000). Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews*, 31(2–3), 357–363.
- Frith, C., & Done, D. (1989). Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action. *Psychological Medicine*, 19, 359–363.
- Gallagher, S. (2004). Neurocognitive models of schizophrenia: A neurophenomenological critique. *Psychopathology*, 37(1), 8–19.
- Garwood, L., Dodgson, G., Bruce, V., & McCarthy-Jones, S. (2013). A preliminary investigation into the existence of a hypervigilance subtype of auditory hallucination in people with psychosis. *Behavioural and Cognitive Psychotherapy*, 1–11.
- Gould, L. N. (1948). Verbal hallucinations and activation of vocal musculature. *American Journal of Psychiatry*, 105, 367–372.
- Gould, L. N. (1950). Verbal hallucinations as automatic speech – The reactivation of dormant speech habit. *American Journal of Psychiatry*, 107(2), 110–119.
- Helmholtz, H. von (1866). Concerning the perceptions in general. In *Treatise on physiological optics* (J. P. C. Southall, Trans., 1925 Opt. Soc. Am. Section 26, reprinted New York: Dover, 1962) (Vol. III, 3rd ed.).
- Heslop, K. (2012). *Binocular rivalry and visuospatial ability in individuals with schizophrenia*. PhD thesis, Queensland University of Technology.
- Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3(96).
- Hohwy, J. (2013). *The Predictive Mind*. OUP.
- Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108(3), 687–701.
- Hong, L., Avila, M., & Thaker, G. (2005). Response to unexpected target changes during sustained visual tracking in schizophrenic patients. *Experimental Brain Research*, 165, 125–131. <http://dx.doi.org/10.1007/s00221-005-2276-z>.
- Jacobsen, E. (1931). Electrical measurements of neuromuscular states during mental activities. VII: Imagination, recollection, and abstract thinking involving the speech musculature. *American Journal of Physiology*, 97, 200–209.
- Johns, L. C., Rossell, S., Frith, C. D., Ahmad, F., Hemsley, D., Kuipers, E., et al (2001). Verbal self-monitoring and auditory verbal hallucinations in patients with schizophrenia. *Psychological Medicine*, 31, 705–715.
- Jones, S. R. (2010). Do we need multiple models of auditory verbal hallucinations? Examining the phenomenological fit of cognitive and neurological models. *Schizophrenia Bulletin*, 36(3), 566–575.
- Jones, S. R., & Fernyhough, C. (2007). Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations. *Consciousness and Cognition*, 16, 391–399.
- Larøi, F. (2012). How do auditory verbal hallucinations in patients differ from those in non-patients? *Frontiers in Human Neuroscience*, 6, 25.
- Larøi, F., Sommer, I. E., Blom, J. D., Fernyhough, C., ffytche, D. H., Hugdahl, K., et al (2012). The characteristic features of auditory verbal hallucinations in clinical and nonclinical groups: State-of-the-art overview and future directions. *Schizophrenia Bulletin*, 38(4), 724–733.
- McCarthy-Jones, S. R., & Fernyhough, C. (2011). The varieties of inner speech: Links between quality of inner speech and psychopathological variables in a sample of young adults. *Consciousness and Cognition*, 20, 1586–1593.
- McCarthy-Jones, S., Trauer, T., Mackinnon, A., Sims, E., Thomas, N., & Copolov, D. L. (2012). A new phenomenological survey of auditory hallucinations: Evidence for subtypes and implications for theory and practice. *Schizophrenia Bulletin*.
- McGuigan, F. (1966). Covert oral behaviour and auditory hallucinations. *Psychophysiology*, 3, 73–80.
- Moseley, P., & Wilkinson, S. (2014). Inner speech is not so simple: A commentary on Cho and Wu (2013). *Frontiers in Psychiatry*, 5, 42. <http://dx.doi.org/10.3389/fpsy.2014.00042>.
- Nayani, T., & David, A. (1996). The auditory hallucination: A phenomenological survey. *Psychological Medicine*, 26(1), 177–189.
- Pearl, D., Yodashtkin-Porat, D., Katz, N., Valevski, A., Aizenberg, D., Sigler, M., et al (2009). Differences in audiovisual integration, as measured by McGurk phenomenon, among adult and adolescent patients with schizophrenia and age-matched healthy control groups. *Comprehensive Psychiatry*, 50, 186–192. <http://dx.doi.org/10.1016/j.comppsy.2008.06.004>.
- Pezzulo, G. (2013). Why do you fear the Bogeyman? An embodied predictive coding model of perceptual inference. *Cognitive, Affective, and Behavioral Neuroscience*.
- Pickering, M., & Clark, A. (2014). Getting ahead: Forward models and their place in cognitive architecture. *Trends in Cognitive Sciences*, 18.
- Ross, L. A., Saint-Amour, D., Leavitt, V. M., Molholm, S., Javitt, D. C., & Foxe, J. J. (2007). Impaired multisensory processing in schizophrenia: Deficits in the visual enhancement of speech comprehension under noisy environmental conditions. *Schizophrenia Research*, 97, 173–183.
- Seal, M. L., Aleman, A., & McGuire, P. K. (2004). Compelling imagery, unanticipated speech and deceptive memory: Neurocognitive models of auditory verbal hallucinations in schizophrenia. *Cognitive Neuropsychiatry*, 9(1–2), 43–72.
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11), 656–663.
- Schneider, U., Leweke, F. M., Sternemann, U., Weber, M. M., & Emrich, H. M. (1996). Visual 3D illusion: A systems-theoretical approach to psychosis. *European Archives of Psychiatry and Clinical Neuroscience*, 246, 256–260.
- Schneider, U., Borsutzky, M., Seifert, J., Leweke, F. M., Huber, T. J., Rollnik, J. D., et al (2002). Reduced binocular depth inversion in schizophrenic patients. *Schizophrenia Research*, 53, 101–108. [http://dx.doi.org/10.1016/S0920-9964\(00\)00172-9](http://dx.doi.org/10.1016/S0920-9964(00)00172-9).
- Schnider, A., & Ptak, R. (1999). Spontaneous confabulators fail to suppress currently irrelevant memory traces. *Nature Neuroscience*, 2(7), 677–681.
- Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, 43(6), 482–489.
- Stephens, G. L., & Graham, G. (2000). *When self-consciousness breaks: Alien voices and inserted thoughts*. Cambridge MA: MIT Press.
- Thaker, G., Ross, D., Cassidy, S., Adami, H., LaPorte, D., et al (1998). Smooth pursuit eye movements to extraretinal motion signals: Deficits in relatives of patients with schizophrenia. *Archives of General Psychiatry*, 55, 830–836.
- Tribus, M. (1961). *Thermodynamics and thermostatics: an introduction to energy, information and states of matter, with engineering applications*. D. Van Nostrand.



- Vygotsky, L. (1934). *Thought and language*. 1934/1986 Cambridge, MA: MIT Press.
- Waters, F., Badcock, J. C., Michie, P. T., & Maybery, M. T. (2006). Auditory hallucinations in schizophrenia: Intrusive thoughts and forgotten memories. *Cognitive Neuropsychiatry*, 11, 65–83.
- Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41, 989–994.
- Winsler, A. (2004). Still talking to ourselves after all these years: Vygotsky, private speech, and self-regulation. *Invited address given at first international symposium on self-regulatory functions of language*, Madrid, November 2004.
- Wolpert, D. (1997). Computational approaches to motor control. *Trends in Cognitive Sciences*, 1(6), 209–216.
- Wu, W. (2012). Explaining schizophrenia: Auditory verbal hallucination and self-monitoring. *Mind and Language*, 27(1), 86–107.